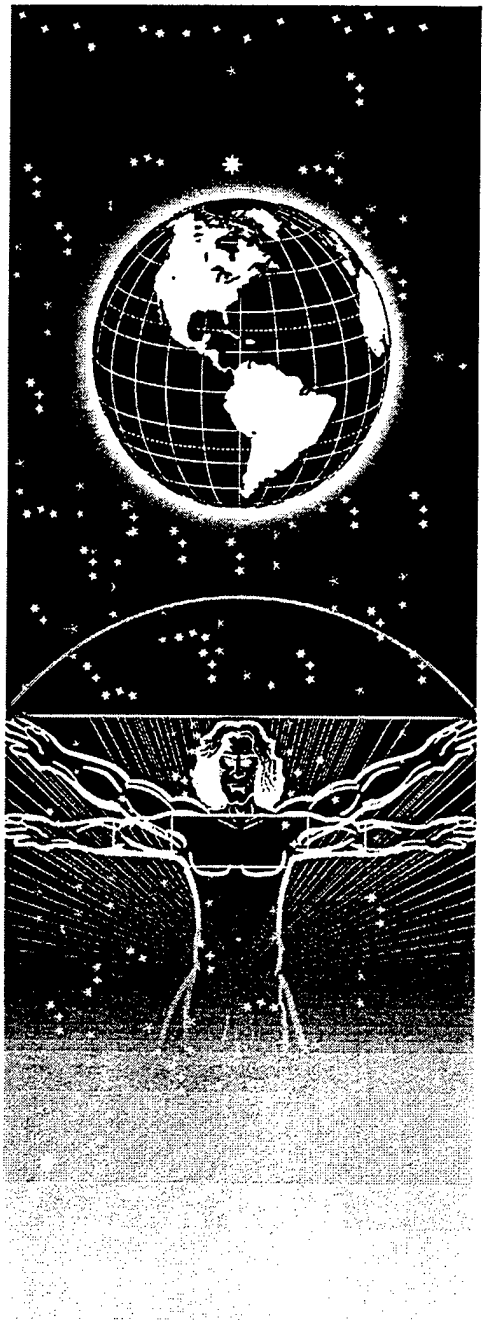
AFRL-HE-AZ-TP-1998-0073



UNITED STATES AIR FORCE RESEARCH LABORATORY

RETEST PERFORMANCE ON AN EXPERIMENTAL COMPUTER-BASED PILOT APTITUDE TEST BATTERY

Thomas R. Carretta

Training Effectiveness Branch
Warfighter Training Research Division
7909 Lindbergh Drive
Brooks AFB TX 78235-5352

September 1998

AIR FORCE MATERIEL COMMAND
AIR FORCE RESEARCH LABORATORY
HUMAN EFFECTIVENESS DIRECTORATE
WARFIGHTER TRAINING RESEARCH DIVISION
6001 South Power Road, Building 558
Mesa AZ 85206-0904

NOTICES

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO exchange).

Using Government drawings, specifications, or other data included in this document for any purpose other than Government-related procurement does not in any way obligate the US Government. The fact that the Government formulated or supplied the drawings, specifications, or other data, does not license the holder or any other person or corporation, or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

THOMAS R. CARRETTA
Project Scientist

DEE H. ANDREWS
Technical Director

LYNN A. CARROLL, Colonel, USAF
Chief, Warfighter Training Research Division

Please notify AFRL/HEOP, 2509 Kennedy Drive, Bldg 125, Brooks AFB, TX 78235-5118, if your address changes, or if you no longer want to receive our technical reports. You may write or call the STINFO Office at DSN 240-3877 or commercial (210) 536-3877, or e-mail Shirley.Walker@platinum.brooks.af.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1998	Final - April 1996 to January 1998

4. TITLE AND SUBTITLE

Retest Performance on an Experimental Computer-Based Pilot Aptitude Test Battery

5. FUNDING NUMBERS

PE - 62205F
PR - 1123
TA - B1
WU - 03

6. AUTHOR(S)

Thomas R. Carretta

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Div, Training Effectiveness Branch
7909 Lindbergh Drive
Brooks AFB TX 78235-5352

8. PERFORMING ORGANIZATION

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Division
6001 South Power Road, Bldg 558
Mesa AZ 85206-0904

10. SPONSORING/MONITORING

AFRL-HE-AZ-TP-1998-0073

11. SUPPLEMENTARY NOTES

Air Force Research Laboratory Technical Monitor: Dr Thomas R. Carretta, (210) 536-3956

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Basic Attributes Test (BAT) is a computer-based battery that contributes to a US Air Force pilot selection composite known as the Pilot Candidate Selection Method (PCSM). When the BAT was operationally implemented in 1993, no retests were allowed because its retest characteristics had not been adequately examined at that point. Results from a recent study of BAT retest characteristics (mean score change, reliability) concluded that a BAT retest could be permitted no less than six months after initial testing. The current research effort examined retest performance on several experimental computer-based tests that are being considered as candidate replacement tests for the BAT. Participants were 340 paid volunteers who completed the test battery and were retested after two weeks, one month, three months, or six months. Results were consistent with those observed earlier for the BAT. First, the experimental computer-based tests demonstrated acceptable reliability. Second, most students showed improvement on retest, regardless of the length of the retest interval. Third, in general, practice effects diminished as the length of the retest interval increased. However, the size and duration of the practice effect varied by test content. A few tests showed performance improvements for the shorter retest intervals (two weeks or one month),but little or no improvements for longer retest intervals (three or six months). Other tests showed moderate to large improvements in performance on retest, even for the six-month retest interval. These results suggest that as with the BAT, retests could be performed no sooner than six months after initial testing.

14. SUBJECT TERMS

Aptitude measurement; Basic Attributes Test; BAT; Computer-based testing; PCSM; Personnel selection; Pilot Candidate Selection Method; Pilot selection; Test-retest performance;

15. NUMBER OF PAGES

26

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

CONTENTS

	Page
INTRODUCTION	1
METHOD	1
Participants	1
Measures	2
Procedure	3
Analyses	3
RESULTS	4
DISCUSSION	8
CONCLUSIONS AND RECOMMENDATIONS	9
REFERENCES	10
APPENDIX A: DETAILED SUMMARY OF TEST-RETEST PERFORMANCE BY RETEST GROUP	11

TABLES

Table No.		
1	Brief Summary of *d* and t-tests by Retest Group	5
A-1	Test-Retest Scores: Two-Week Retest Group (N = 54)	12
A-2	Test-Retest Scores: One-Month Retest Group (N = 188)	14
A-3	Test-Retest Scores: Three-Month Retest Group (N = 51)	16
A-4	Test-Retest Scores: Six-Month Retest Group (N = 47)	18

PREFACE

This research was conducted under Work Unit 1123-B1-03, Practice Effects II, in support of aircrew selection and classification research and development (R&D). Work unit monitor was Dr Thomas R. Carretta.

The objective of this effort was to estimate the change in scores on computer-based ability tests as a result of repeat testings. The participants were paid volunteers from the San Antonio, TX, area. The data were collected by Carol Maske (Metrica, Inc.) under Contract F41624-95-D-5030.

RETEST PERFORMANCE ON AN EXPERIMENTAL COMPUTER-BASED PILOT APTITUDE TEST BATTERY

INTRODUCTION

The Basic Attributes Test (BAT) is a computer-based test that contributes to a US Air Force pilot selection composite known as the Pilot Candidate Selection Method or PCSM (Carretta, 1992). When the BAT was operationally implemented for pilot selection in 1993, the US Air Force allowed no retests (see Air Force Instruction 36-2605, 17 June 94). A recent study (Carretta, Zelenski, & Ree, 1997) examined retest reliability and mean score change for the BAT. After 477 college students completed the BAT and retested in two weeks, three months, or six months, several important results were observed. First, BAT scores showed acceptable retest reliability. Second, scores for about 70% of the college students improved on retest, regardless of length of retest interval. Those who performed poorly on the first test generally exhibited larger improvements than those who performed well on the first test. Third, practice effects diminished as the length of the retest interval increased. For a six-month retest interval, it is expected that PCSM scores would increase on average by about six percentile points. It was concluded that BAT retests could be permitted no less than six months after initial testing. Based on these results, the US Air Force decided to allow pilot applicants one retest on BAT after at least a six-month interval. This retest policy is consistent to that already in practice for another US Air Force personnel selection test, the Air Force Officer Qualifying Test (AFOQT) (Carretta & Ree, 1997).

Over the past several years, the US Air Force has conducted basic and exploratory studies to develop several new computer-based tests that are being considered as candidate replacement tests for the BAT (see Carretta, 1996; Carretta, Perry, & Ree, 1996). The objective of the current study was to examine mean score performance and reliability for some of these experimental tests in the event of a retest. These data could be used to inform policy makers regarding the expected changes to mean test scores and rank ordering of retesters in the event of a retest.

METHOD

Participants

Participants were 340 paid volunteers from the San Antonio, TX, area. Although some were recruited through a temporary employment agency, most participants were recruited through advertisements targeted toward college student populations. The sample had almost equal numbers of men (48.9%) and women (51.1%). Most participants identified themselves as either Caucasian (47.3%) or Hispanic (39.2%), with relatively small numbers of African Americans (6.5%), Asians (3.7%), or other/unidentified (3.3%). Their ages at time of the first test ranged from 18 to 32 years with a mean of 23.1 years. Participants were informed that the study involved the evaluation of several experimental US Air Force personnel measurement tests.

Measures

Anticipation. In this velocity estimation test, a target moves horizontally across the screen from left to right. When the target reaches line "A" on the screen, it disappears from view but continues to move at the same velocity. The participant's task is to estimate when the target will cross line "B" (to the right of line "A"). Target velocity and point of disappearance vary across test items. Scores include average distance error and average response time error.

Laser Aiming Task 1. This test (Tirre & Raouf, 1994) measures the psychomotor factors of multilimb coordination and aiming. Participants maneuver left and right foot pedals to aim a "laser gun" at aircraft moving horizontally across the screen. Participants fire the "laser gun" by pressing the ENABLE key. Speed, distance, and direction (left or right) of the target aircraft vary across trials. Scores include the number of shots fired, the number of hits, and area on the screen where the target was hit (x-range).

Laser Aiming Task 2. As with Laser Aiming Task 1, this test (Tirre & Raouf, 1994) assesses multilimb coordination and aiming. It is similar to Laser Aiming Task 1, except that participants are instructed to imagine they are shooting from an aircraft located at the bottom of the screen. Participants must match the apparent altitude (size) of the target and the "laser gun" to get the laser beam on target. Scoring is similar to Laser Aiming Task 1.

Matrices. This spatial reasoning test is similar to Raven's (1966) Matrices. Participants are shown an incomplete geometric pattern (the lower right hand corner is missing) and must choose from several alternatives, which would correctly complete the pattern. Scores include average response time and percent correct.

Pitch-Roll-Yaw. In this spatial visualization test, representations of two aircraft are displayed side-by-side (Tirre & Raouf, 1994). The aircraft on the left is a stationary target. Participants must use the right-hand control stick and the rudder pedals to maneuver the aircraft on the right to match the stationary aircraft on the left (target) in the pitch, roll, and yaw axes. A "match" occurs when the participant maneuvers the aircraft within a 5-deg tolerance. Although test instructions emphasize accuracy rather than speed, both accuracy and solution time are recorded.

Rapid Serial Classification: 4-Square. This test measures spatial reasoning ability. Participants are shown a 4-square (2-by-2) display in which a letter pattern can be drawn (C, X, or Z) between points. Participants must determine which letter is being drawn by following the pattern of dots as they are sequentially illuminated and extinguished. Average response time and percent correct are scored.

Scheduling 2. In this divided attention test, five horizontal logarithmic scales are presented. A line beneath each scale increases at a unique, constant rate. Each line and scale appears on a separate screen that may be viewed by entering the scale number on the response keypad. Participants score points equal to the current value of the line displayed on the scale by pressing the ENABLE key. When the ENABLE key is pressed, the participant's total score is

incremented by the value of the line which is then reset to 0, where it will start increasing again. If the value of a line reaches the upper limit of the scale, and the participant has not responded by pressing the ENABLE key, the value of the line will reset to 0 without the participant receiving any points. Scores include the total number of points accumulated and the ratio of total points accumulated divided by total points possible.

Synthesis Add & Subtract. This test measures spatial working memory. The task requires combining or deleting simple line figures assigned to three letters (X, Y, and Z). Two figures are assigned to each letter in the form of an addition or subtraction equation. Participants must mentally combine or delete the lines of these figures and then memorize the combination. Information about one figure is sometimes needed to solve the equation for one of the other figures. Scores include average response time and percent correct.

Time Sharing 2. This test provides measures of attention and the psychomotor factors of reaction time and rate control (Fleishman, 1964). The first part of the test involves learning a compensatory tracking task, where participants maneuver the right-hand control stick to keep a "gunsight" centered on an airplane. The second part of the test involves learning an attention task. Numbers appear one at a time in sequence at the lower part of the screen. The number sequence is 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0, 1, 2, 3, etc. Occasionally, a number will be missing from the sequence (e.g., 0, 1, 2, 3, 5, 6, 7, ... [4 is missing]). Participants are required to type the missing number on the keypad. During the final part of this test, participants simultaneously perform tracking and attention tasks. Scores include tracking performance and accuracy in responding to missing numbers.

Procedure

Each participant completed the test battery and was assigned randomly to one of four retest intervals: 2 weeks (n = 54), 1 month (n = 188), 3 months (n = 51), or 6 months (n = 47). Each participant retested on the test battery at the completion of one of the retest intervals. No practice was permitted between the first and second test.

Analyses

Analyses consisted of examination of mean score changes on retest and the correlation between the first and second test scores.

Mean scores. Differences between first and second administration means were expressed in standard deviation units or *d* (i.e., $[\overline{X}_1 - \overline{X}_2]/ S_D$). The standard deviation for *d* was defined as the within-group standard deviation calculated from the weighted average of the square root of the variances for the scores being compared (e.g., first versus second test for the 2 week interval group). *d* values frequently are used as an estimate of effect size. Others (e.g., Cohen, 1988) interpret *d* values of .20 as "small," .50 "medium," and .80 "large". In addition to the

computation of *d*, one-tailed paired-samples t-tests were performed to examine whether performance *improved* on retest. A .01 Type I error rate was used for the t-tests.

Note that improvements in tracking distance error (Anticipation) and response time (Anticipation, Matrices, Pitch-Roll-Yaw, Rapid Serial Classification, Synthesis Add & Subtract, and Time Sharing 2) will result in positive values for the *d* and t-tests (i.e., second time means should be lower reflecting smaller errors). Improvements in percentage scores (Laser Aiming Task 1, Laser Aiming Task 2, Matrices, Pitch-Roll-Yaw, Rapid Serial Classification, Scheduling 2, Synthesis Add & Subtract, and Time Sharing 2) will result in negative values for *d* and t-tests (i.e., second time means should be higher reflecting greater accuracy or efficiency).

Test-Retest Correlations. Correlations between first and second test scores indicate the extent to which the rank order of participants on the first test change after retesting (i.e., is the ranking on the second test the same as the ranking on the first test?). Test-retest correlations also provide an estimate of reliability.

RESULTS

Results varied by test. Table 1 provides a brief summary of the mean score analyses and test-retest correlations for the two-week, one-month, three-month, and six-month retest groups. More detailed results including the means and standard deviations of the test scores for each retest group are provided in the Appendix in Tables A-1 through A-4.

Anticipation. The carry-over benefit from retesting (i.e., reduction in time and distance error) was short-lived for the Anticipation test. Only the two-week retest group showed significant mean score improvement in performance on retest. The average *d* value across the two error scores (response time and distance) were 0.42, -0.04, 0.06, and 0.17 for the two-week, one-month, three-month, and six-month retest groups.

The correlations between first and second test scores indicated moderate agreement in rank order between first and second tests and somewhat lower than desirable retest reliability. The average test-retest reliabilities for the four groups were .677, .722, .612, and .690.

Laser Aiming Tasks 1 and 2. Performance improved on retest for both psychomotor aiming tests (Laser Aiming Task 1 and Laser Aiming Task 2). Generally, on retest, participants needed fewer shots to hit the targets, had more "hits," and hit the targets earlier in their flight paths (higher x-range score). Although the amount of score improvement decreased as the length of the retest interval increased, small to moderate mean score improvements in performance were observed for those tested after a six-month retest interval. For Laser Aiming Task 1, the average *d* across all three subscores (N shots fired, N hits, and x-range) was 0.84, 0.68, 0.81, and 0.29 respectively for the two-week, one-month, three-month, and six-month retest groups. The average *d* values for Laser Aiming Task 2 were 1.74, 0.57, 0.46, and 0.58. The mean score changes were somewhat *greater* for the Laser Aiming Tasks than were observed for the operational BAT psychomotor composite (d = 0.48, 0.33, and 0.25 for two-week, three-month, and six-month retest intervals; see Carretta et al., 1977).

Table 1. Brief Summary of *d* and t-tests by Retest Group

	Two-week (N = 54)		One-month (N = 188)		Three-month (N = 51)		Six-month (N = 47)	
Test Score	*d*	r_{12}	*d*	r_{12}	*d*	r_{12}	*d*	r_{12}
Anticipation								
Response Time Error	0.39*	.674	-0.06	.710	-0.03	.561	0.15	.659
Distance Error	0.45*	.679	-0.02	.733	0.15	.662	0.18	.721
Laser Aiming Task 1								
N Shots Fired	0.81*	.713	0.70*	.577	0.92*	.660	-0.05	.473
N Hits	-0.95*	.485	-0.87*	.573	-0.77*	.516	-0.34	.446
Avg. X-Range	-0.76*	.760	-0.46*	.608	-0.45*	.702	-0.59*	.556
Laser Aiming Task 2								
N Shots Fired	0.99*	.726	0.47*	.655	0.35*	.649	0.29	.637
N Hits	-1.65*	.829	-0.93*	.727	-0.87*	.707	-0.99*	.673
Avg. X-Range	-2.57*	.535	-0.32*	.601	-0.17	.425	-0.45*	.637
Matrices								
Avg. Response Time	0.47*	.772	0.43*	.631	0.15	.688	-0.07	.692
% Correct	-0.39*	.735	-0.14	.661	-0.28	.623	-0.07	.716
Pitch-Roll-Yaw								
Avg. Response Time	-0.21	.636	0.04	.609	0.08	.435	0.01	.693
% Correct	-0.32	.798	-0.30*	.654	-0.08	.579	-0.42*	.812
Rapid Serial Classification: 4-Square								
Avg. Response Time	0.57*	.689	0.61*	.736	0.39*	.567	0.28	.514
% Correct	-0.79*	.856	-0.83*	.863	-0.35*	.817	-0.85*	.859
Scheduling 2								
Points Achieved (PA)	-1.19*	.862	-0.78*	.756	-0.58*	.759	-0.92*	.818
Points Possible (PP)	0.35*	.218	-0.07	.451	0.03	.395	0.19	.551
Ratio (PA/PP)	-0.79*	.856	-0.59*	.718	-0.46*	.754	-0.83*	.799

Table 1. Brief Summary of *d* and t-tests by Retest Group (Cont'd.)

	Two-week (N = 54)		One-month (N = 188)		Three-month (N = 51)		Six-month (N = 47)	
Test Score	*d*	r_{12}	*d*	r_{12}	*d*	r_{12}	*d*	r_{12}
Synthesis Add & Subtract								
Avg. Response Time	0.57*	.600	0.56*	.654	0.50*	.624	0.60*	.630
% Correct	-0.59*	.726	-0.35*	.678	-0.42*	.783	-0.56*	.783
Time Sharing 2								
Avg. Response Time	0.59*	.754	0.35*	.682	0.45*	.596	0.68*	.721
% Correct	-0.42*	.596	-0.19*	.602	-0.34*	.531	-0.23	.385

Notes. All t-tests were one-tailed. r_{12} is the correlation between the first and second test score
*$p < .01$ (critical t value varies by sample size).

The correlations between first and second test scores indicated low to moderate agreement in rank order between first and second tests and less than acceptable test-retest reliability. Generally, as the length of the retest interval increased, retest reliability decreased. For Laser Aiming Task 1, the average correlations across all three subscores (number of shots fired, number of hits, and x-range) were .653, .586, .626, and .492 for the two-week, one-month, three-month, and six-month groups. The retest reliabilities for Laser Aiming Task 2 were slightly higher and indicated more stability in rank order across retest interval: .697, .664, .594, and .649. These test-retest reliabilities were lower than observed for the BAT psychomotor composite: .800, .801, and .775 for the two-week, three-month, and six-month groups (Carretta et al., 1977).

Matrices. Although both response time and accuracy are recorded on this test, accuracy (i.e., percent correct) is more important for scoring purposes. This is also true for Pitch-Roll-Yaw, Rapid Serial Classification: 4-Square, and Synthesis Add & Subtract.

As with Anticipation, retest performance improvements based on previous exposure to the test diminished as the length of the retest interval increased. The average *d* values for the response time scores were 0.47, 0.43, 0.15, and -0.07. The average *d* values for the percent correct score were -0.39, -0.14, -0.28, and -0.07. The retest reliabilities were moderate, but acceptable for measures of this type.

Pitch-Roll-Yaw. Results for the Pitch-Roll-Yaw Test were mixed. There was very little improvement in response time on retest (*d* = -0.21, 0.04, 0.08, and 0.01). This was expected as the test instructions emphasize accuracy, not speed. Larger improvements in performance were observed for the percent correct score (*d* = -0.32, -0.30, -0.08, and -0.42).

As with Matrices, the retest reliabilities were moderate, but acceptable. Improvements could be made by increasing the number of items (this version had only 12 items) or by changing the scoring. Currently, items are scored dichotomously as correct/incorrect. A more appropriate scoring strategy might be to compute the amount of angular displacement along the pitch, roll, and yaw axes.

Rapid Serial Classification: 4-Square. Retesting results showed consistent moderate to large increases in performance (i.e., quicker response time, greater accuracy). As with Matrices and Pitch-Roll-Yaw, although response time is recorded on this test, accuracy is more important to the scoring. For average response time, the *d* values were 0.57, 0.61, 0.39, and 0.28. The *d* values for response accuracy were -0.79, -0.83, -0.35, and -0.85.

Retest reliabilities were acceptable. Reliabilities for the accuracy score were .817 or greater.

Scheduling 2. The most important scores from this test are the total number of points achieved and the ratio of points achieved/points possible (i.e., scoring efficiency). Participants showed moderate to large mean improvements on both of these scores, even in the six-month retest group. The *d* values for points achieved were -1.19, -0.78, -0.58, and -0.92. The *d* values for the ratio of points achieved/points possible were -0.79, -0.59, -0.46, and -0.83.

Retest reliabilities were acceptable for both points achieved (.862, .756, .759, and .818) and the ratio score (.856, .718, .754, and .799).

Synthesis Add & Subtract. Participants showed moderate improvement in performance on retests (i.e., quicker response time, greater accuracy), regardless of the length of the retest interval. The *d* values for average response time were 0.57, 0.56, 0.50, and 0.60. The *d* values for percent correct were -0.59, -0.35, -0.42, and -0.56.

Retest reliabilities were slightly low for response time (.600, .654, .624, and .630), but were acceptable for response accuracy (.726, .678, .783, and .783).

Time Sharing 2. For Time Sharing 2, there is no clear relationship between length of retest interval and performance on retest, especially for the response time score. The average response time scores all indicated moderate improvement in performance (i.e., quicker response time) on retest (d = 0.59, 0.56, 0.45, and 0.68). The greatest mean score improvement on retest in accuracy occurred for the two-week retest group (d = -0.42). The amount of improvement was less for the one-month (-0.19), three-month (-0.34), and six-month (-0.23) groups.

Retest reliabilities fluctuated across retest intervals, showing no consistent pattern. Retest reliability was low for the percent correct score, especially for the six-month retest group (.596, .602, .531, and .385). The low reliability for the accuracy score may be due to the dichotomous response format of the "missing digit" task.

DISCUSSION

Retesting results varied by test and, sometimes, by scores within a test. As expected, in general the size of mean score improvements decreased as the length of the retest interval increased (e.g., Anticipation, Matrices). Notable exceptions included the Scheduling 2 and Syntheses Add & Subtract tests where the mean score improvements were almost equal in size for all retest groups. We will focus on the results from the six-month retest group since the operational US Air Force policy requires *at least a six-month retest interval* for retesting on the AFOQT and has recently adopted the same retest policy for the BAT.

The mean *d* value for all scores for the six-month retest group was 0.41 (after reflecting negative signs to indicate score improvement where appropriate) and ranged from 0.01 (Pitch-Roll-Yaw, average response time) to -0.99 (Laser Aiming Task 2, number of hits). The mean retest reliability was .656 and ranged from .446 (Laser Aiming Task 1, number of hits) to .859 (Rapid Serial Classification: 4 Square, percent correct). These values are similar to those from recent studies of the retest characteristics of the AFOQT (Carretta & Ree, 1997) and BAT (Carretta et al., 1977).

Carretta and Ree (1997; Table A-2) reported a mean *d* of 0.41 across all 16 AFOQT subtests with a range from 0.28 (General Science) to 0.61 (Instrument Comprehension). Although the *d* values from the current study were more variable than those reported for the AFOQT subtests, the mean *d* values for the two studies were almost identical. The mean and range of reliabilities also were similar for the experimental computer-based tests and the AFOQT. The mean of the AFOQT subtest retest reliabilities was .686 and ranged from .485 (Hidden Figures) to .822 (Word Knowledge).

Carretta, et al. (1977, Table 3) examined retest performance on the BAT. The average *d* was a little larger for the experimental tests (0.41) than for the BAT (0.30). Consistent with the AFOQT comparisons, the *d* values from the current study were more variable than those reported for the BAT. For the six-month retest group, the *d* values for the 11 BAT subscores went from 0.01 (Time Sharing, average tracking difficulty) to 0.64 (Activities Interest Inventory, average response time). The mean BAT retest reliability was slightly higher than found for either the current study or the AFOQT. The BAT retest reliabilities had a mean of .719 and ranged from .474 (Time Sharing, average response time) to .856 (Activities Interest Inventory, percent choices).

The test-retest correlations from the experimental computer-based tests suggested that the relative order of the retesters stayed about the same after retesting. However, as with the AFOQT and BAT, mean score performance tended to improve on retesting. As with the AFOQT and BAT, the mean scores of those who retest on these experimental tests will improve relative to those who choose not to retest.

CONCLUSIONS AND RECOMMENDATIONS

These results are generally consistent with previous research. The amount of score improvement for these experimental tests was comparable to that observed for the AFOQT and BAT. Also, the test-retest correlations suggested that the relative order of the retesters stayed the same after retesting. Retest reliability may be improved for some tests by either increasing the number of items or considering other scoring options (e.g., Pitch-Roll-Yaw, angle error instead of correct/incorrect).

In the event that these experimental tests someday replace the operational BAT, these results suggest that a retest could be allowed after at least a six-month test-retest interval. The current Air Force policy is to report only the most recent test score to pilot training selection boards and not indicate whether it represents a first test or a retest. Additional studies of these experimental tests should be done to examine the validity of first versus retest scores against pilot training outcome.

REFERENCES

Carretta, T. R. (1992). Recent developments in U. S. Air Force pilot candidate selection and classification. *Aviation, Space, and Environmental Medicine, 63,* 1112-1114.

Carretta, T. R. (1996). *Preliminary validation of several US Air Force computer-based cognitive pilot selection tests* (AL/HR-TP-1996-0008). Brooks Air Force Base, TX: Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division.

Carretta, T. R., Perry, D. C., Jr., & Ree, M. J. (1996). Prediction of situational awareness in F-15 pilots. *The International Journal of Aviation Psychology, 6,* 21-41.

Carretta, T. R., & Ree, M. J. (1997). *The best retest is the average: Findings and implications.* (AL/HR-TP-1996-0021). Brooks Air Force Base, TX: Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division.

Carretta, T. R., Zelenski, W. E., & Ree, M. J. (1977). Basic Attributes Test (BAT) retest performance. (AL/HR-TP-1997-0040). Brooks Air Force Base, Texas. Armstrong Laboratory, Human Resources Directorate, Aircrew Training Research Division

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fleishman, E. A. (1964). *The structure and measurement of physical fitness*. Englewood Cliffs, NJ: Prentice-Hall.

Raven, J. C. (1966). *Advanced Progressive Matrices*. New York: Psychological Corporation.

Tirre, W. C., & Raouf, K. K. (1994). Gender differences in perceptual-motor performance. *Aviation, Space, and Environmental Medicine, 65*, A49-A53.

APPENDIX A

DETAILED SUMMARY OF TEST-RETEST PERFORMANCE BY RETEST GROUP

Table A-1. Test-Retest Scores: Two-Week Retest Group (N = 54)

Test Score	First Test Average	Second Test Average	S_D	*d*	t	r_{12}
Anticipation						
Response Time Error	569.11	520.64	123.23	0.39	2.86*	.674
Distance Error	34.39	31.41	6.70	0.45	3.24*	.679
Laser Aiming Task 1						
N Shots Fired	245.57	218.26	33.53	0.81	5.93*	.713
N Hits	91.96	99.91	8.37	-0.95	-6.92*	.485
Avg. X-Range	3.21	3.37	0.21	-0.76	-5.55*	.760
Laser Aiming Task 2						
N Shots Fired	505.76	393.37	114.11	0.99	7.17*	.726
N Hits	48.20	67.39	11.65	-1.65	-11.99*	.829
Avg. X-Range	2.65	3.37	0.28	-2.57	-18.72*	.535
Matrices						
Avg. Response Time	16,109.20	14,938.42	2,486.30	0.47	3.43*	.772
% Correct	68.96	72.21	8.42	-0.39	-2.81*	.735
Pitch-Roll-Yaw						
Avg. Response Time	52,740.98	59,038.43	29,952.98	-0.21	-1.53	.636
% Correct	28.70	34.42	17.87	-0.32	-2.33	.798
Rapid Serial Classification:						
4-Square						
Avg. Response Time	1,005.38	935.33	123.86	0.57	4.12*	.689
% Correct	57.04	64.45	9.36	-0.79	-5.76*	.856
Scheduling 2						
Points Achieved (PA)	3,437.44	4,475.48	870.10	-1.19	-8.69*	.862
Points Possible (PP)	9,326.22	8,828.44	1,414.44	0.35	2.56*	.218
Ratio (PA/PP)	57.04	64.45	9.36	-0.79	-5.76*	.856
Synthesis Add & Subtract						
Avg. Response Time	4,813.85	4,175.00	1,131.32	0.57	4.11*	.600
% Correct	59.34	71.69	20.91	-0.59	-4.30*	.726

Table A-1. Test-Retest Scores: Two-Week Retest Group (N = 54) (Cont'd.)

Test Score	First Test Average	Second Test Average	S_D	*d*	t	r_{12}
Time Sharing 2						
Avg. Response Time	1,615.88	1,475.37	239.69	0.59	4.27*	.754
% Correct	79.04	84.37	12.58	-0.42	-3.08*	.596

Notes. All t-tests were one-tailed. r_{12} is the correlation between the first and second test score. S_D is the within-group standard deviation of the difference between the first and second test administrations.
*$p < .01$ (critical t value for 53 df = 2.399)

Table A-2. Test-Retest Scores: One-Month Retest Group (N = 188)

Test Score	First Test Average	Second Test Average	S_D	*d*	t	r_{12}
Anticipation						
Response Time Error	570.94	582.62	186.48	-0.06	-0.86	.710
Distance Error	34.83	35.01	10.61	-0.02	-0.23	.733
Laser Aiming Task 1						
N Shots Fired	238.70	209.75	41.50	0.70	9.54*	.577
N Hits	93.93	100.11	7.12	-0.87	11.87*	.573
Avg. X-Range	3.20	3.31	0.24	-0.46	-6.27*	.608
Laser Aiming Task 2						
N Shots Fired	436.69	379.98	119.65	0.47	6.48*	.655
N Hits	50.14	64.29	15.24	-0.93	-12.70*	.727
Avg. X-Range	2.57	2.63	0.19	-0.32	-4.32*	.601
Matrices						
Avg. Response Time	16,196.69	14,613.21	3,693.73	0.43	5.86*	.631
% Correct	70.05	71.71	12.20	-0.14	-1.86	.661
Pitch-Roll-Yaw						
Avg. Response Time	53,975.86	52,983.44	24,451.03	0.04	0.56	.609
% Correct	26.68	33.46	22.28	-0.30	-4.16*	.654
Rapid Serial Classification:						
4-Square						
Avg. Response Time	1,014.55	939.82	122.82	0.61	8.32*	.736
% Correct	57.87	66.18	10.05	-0.83	-11.31*	.863
Scheduling 2						
Points Achieved (PA)	3,773.39	4,617.83	1,085.11	-0.78	-10.64*	.756
Points Possible (PP)	9,408.34	9,619.74	2,925.41	-0.07	-0.07	.451
Ratio (PA/PP)	42.50	51.15	14.57	-0.59	-8.12*	.718
Synthesis Add & Subtract						
Avg. Response Time	4,887.99	4,154.74	1,317.22	0.56	7.61*	.654
% Correct	64.63	71.82	20.69	-0.35	-4.75*	.678

Table A-2. Test-Retest Scores: One-Month Retest Group (N = 188) (Cont'd.)

Test Score	First Test Average	Second Test Average	S_D	*d*	t	r_{12}
Time Sharing 2						
Avg. Response Time	1,609.48	1,502.07	307.69	0.35	4.77*	.682
% Correct	78.68	81.62	15.56	-0.19	-2.58*	.602

Notes. All t-tests were one-tailed. r_{12} is the correlation between the first and second test score. S_D is the within-group standard deviation of the difference between the first and second test administrations

*$p < .01$ (critical t value for 187 df = 2.346)

Table A-3. Test-Retest Scores: Three-Month Retest Group (N = 51)

Test Score	First Test Average	Second Test Average	S_D	*d*	t	r_{12}
Anticipation						
Response Time Error	518.70	522.68	132.27	-0.03	-0.21	.561
Distance Error	33.17	32.27	6.13	0.15	1.04	.662
Laser Aiming Task 1						
N Shots Fired	252.51	222.82	32.28	0.92	6.50*	.660
N Hits	92.12	98.29	8.03	-0.77	-5.43*	.516
Avg. X-Range	3.17	3.26	0.20	-0.45	-3.18*	.702
Laser Aiming Task 2						
N Shots Fired	463.90	424.47	113.24	0.35	2.46*	.649
N Hits	48.08	60.31	14.12	-0.87	-6.13*	.707
Avg. X-Range	2.59	2.63	0.24	-0.17	-1.18	.425
Matrices						
Avg. Response Time	15,536.59	15,108.73	2,769.54	0.15	1.09	.688
% Correct	72.25	75.08	9.85	-0.28	2.03	.623
Pitch-Roll-Yaw						
Avg. Response Time	59,464.00	56,632.63	34,192.78	0.08	0.59	.435
% Correct	28.44	30.06	21.73	-0.08	-0.53	.579
Rapid Serial Classification: 4-Square						
Avg. Response Time	954.67	898.84	144.32	0.39	2.74*	.567
% Correct	58.98	62.99	11.37	-0.35	-2.49*	.817
Scheduling 2						
Points Achieved (PA)	4,090.04	4,664.33	982.81	-0.58	-4.13*	.759
Points Possible (PP)	8,795.78	9,728.00	2,212.54	0.03	0.22	.395
Ratio (PA/PP)	46.50	52.24	12.62	-0.46	-3.21*	.754
Synthesis Add & Subtract						
Avg. Response Time	4,147.82	3,715.88	869.61	0.50	3.51*	.624
% Correct	71.18	77.50	14.97	-0.42	2.99*	.783

Table A-3. Test-Retest Scores: Three-Month Retest Group (N = 51) (Cont'd.)

Test Score	First Test Average	Second Test Average	S_D	*d*	t	r_{12}
Time Sharing 2						
Avg. Response Time	1,523.44	1,390.98	296.51	0.45	3.16*	.596
% Correct	79.76	84.39	13.61	-0.34	-2.41*	.531

Notes. All t-tests were one-tailed. r_{12} is the correlation between the first and second test score. S_D is the within-group standard deviation of the difference between the first and second test administrations.

*$p < .01$ (critical t value for 50 df = 2.403)

Table A-4. Test-Retest Scores: Six-Month Retest Group (N = 47)

Test Score	First Test Average	Second Test Average	S_D	d	t	r_{12}
Anticipation						
Response Time Error	523.34	502.77	133.30	0.15	1.05	.659
Distance Error	31.99	30.83	6.60	0.18	1.19	.721
Laser Aiming Task 1						
N Shots Fired	229.43	231.51	44.48	-0.05	-0.32	.473
N Hits	94.55	97.34	8.24	-0.34	-2.30	.446
Avg. X-Range	3.20	3.33	0.22	-0.59	-4.01*	.556
Laser Aiming Task 2						
N Shots Fired	430.79	396.06	120.52	0.29	1.95	.637
N Hits	50.49	65.15	14.83	-0.99	-6.71*	.673
Avg. X-Range	2.56	2.65	0.20	-0.45	-3.05*	.637
Matrices						
Avg. Response Time	15,456.36	15,643.64	2,886.81	-0.07	-0.44	.692
% Correct	72.18	72.71	8.13	-0.07	-0.44	.716
Pitch-Roll-Yaw						
Avg. Response Time	59,782.79	59,561.26	26,436.61	0.01	0.06	.693
% Correct	31.03	38.30	17.26	-0.42	-2.86*	.812
Rapid Serial Classification:						
4-Square						
Avg. Response Time	967.38	928.22	141.88	0.28	1.87	.514
% Correct	61.26	69.30	9.44	-0.85	-5.78*	.859
Scheduling 2						
Points Achieved (PA)	3,996.79	4,750.36	819.59	-0.92	-6.24*	.818
Points Possible (PP)	9,277.28	9,038.98	1,249.00	0.19	1.29	.551
Ratio (PA/PP)	44.74	53.71	10.85	-0.83	-5.61*	.799
Synthesis Add & Subtract						
Avg. Response Time	4,623.53	4,050.68	959.87	0.60	4.05*	.630
% Correct	68.17	76.14	14.33	-0.56	-3.77*	.783

Table A-4. Test-Retest Scores: Six-Month Retest Group (N = 47) (Cont'd.)

Table A-4. Test-Retest Scores: Six-Month Retest Group (N = 47)

Test Score	First Test Average	Second Test Average	S_D	d	t	r_{12}
Time Sharing 2						
Avg. Response Time	1,520.47	1,379.60	208.05	0.68	4.59*	.721
% Correct	83.49	86.55	13.14	-0.23	1.58	.385

Notes. All t-tests were one-tailed. r_{12} is the correlation between the first and second test score. S_D is the within-group standard deviation of the difference between the first and second test administrations.

*$p < .01$ (critical t value for 46 df = 2.410)